

De la cooccurrence généralisée à la variation du sens lexical

Matthias Tauveron



Édition électronique

URL : <http://journals.openedition.org/corpus/2236>

DOI : 10.4000/corpus.2236

ISSN : 1765-3126

Éditeur

Bases ; corpus et langage - UMR 6039

Édition imprimée

Date de publication : 1 janvier 2012

ISSN : 1638-9808

Référence électronique

Matthias Tauveron, « De la cooccurrence généralisée à la variation du sens lexical », *Corpus* [En ligne], 11 | 2012, mis en ligne le 18 juin 2013, consulté le 08 septembre 2020. URL : <http://journals.openedition.org/corpus/2236> ; DOI : <https://doi.org/10.4000/corpus.2236>

© Tous droits réservés

De la cooccurrence généralisée à la variation du sens lexical

Matthias TAUVERON

UR LiLPa, Fonctionnements Discursifs et Traduction,
Université de Strasbourg, 22 rue Descartes, F- 67084 Strasbourg
Cedex, France

1. Introduction : une problématique de sens lexical en contexte

1.1 *Le mot dans le texte*

Cette étude¹ s'intéresse aux relations entre les mots d'un texte et la textualité. Comment un mot s'intègre-t-il à l'ensemble du lexique d'un texte ? Comment le lexique d'un texte est-il organisé ? On sait que, au-delà de l'étude des seules occurrences des mots, l'étude de la cooccurrence est en soi une étude de l'organisation des mots en discours. En synthétisant sur une même vue (cf. Figure 1 *infra*) l'ensemble des relations de cooccurrence entre les mots d'un texte, on accède à une représentation de la textualité qui permet de cerner le rôle de chaque mot dans la construction de l'ensemble du lexique textuel. Inversement, on peut en tirer des faits concernant la place que le texte donne à chacun de ses mots, et sur l'influence que peut avoir le contexte – pris dans sa globalité – sur le sens des mots. Cette étude globale des relations de cooccurrence – appelée par Viprey (2006) *cooccurrence généralisée* – permet d'aborder l'interaction entre le sens des unités lexicales et le sens des textes. « Tout mot placé dans un texte en reçoit des déterminations sémantiques, et modifie potentiellement le signifié de chacun des mots

· matthias.tauveron@etu.unistra.fr

¹ L'auteur tient à remercier ici Damon Mayaffre de lui avoir témoigné si spontanément sa confiance, et, si longuement son intérêt pour ce travail.

qui le composent » (Valette 2010), « dans la plupart des cas, on ne peut pas parler de déductibilité à sens unique : il y a influence réciproque entre les divers éléments » (Victorri & Fuchs 1996 : 174-175).

1.2 Le nom action en traduction

1.2.1 Le sens du nom action en discours

Cette question des influences et des déterminations sémantiques entre les mots d'un même texte est abordée ici par l'étude du nom français *action*². Si ce mot semble *a priori* monosémique³, on n'en a pas moins l'intuition que son sens semble vaste, voire vague, et donc *a priori* susceptible d'avoir une importante variabilité en discours. La présente étude cherche à expliquer des cas de variabilité en discours de son sens, une fois ces cas constatés par des différences de traduction.

En effet, si *action* est traduit en suédois par un nom dans 89 % des cas, il n'a pas d'équivalent unique dans cette langue : au moins deux noms (*handling*, dans 18 % des cas, *åtgärd*, 44 % des cas) le traduisent fréquemment. Nous posons une première hypothèse, selon laquelle cette différence de traduction correspond à une différence de sens⁴. Cependant, un regard à l'œil nu⁵ sur le corpus ne permet pas d'expliquer pourquoi telles occurrences connaissent telle traduction, comme on le voit sur ces deux exemples :

2 Notre corpus est formé de 400 occurrences de ce nom, tirées des débats tenus au Parlement Européen, tels qu'ils figurent dans la base de données Europarl (Koehn, 2005).

3 Les occurrences sont sélectionnées de façon que ce terme soit monosémique. En l'occurrence, nous excluons les quelques cas dans lesquelles c'est son sens financier qui est actualisé.

4 Nous supposons également que l'emploi des mots tel qu'il apparaît dans ce corpus reflète bien globalement son emploi dans le discours et le genre en question (débat politique), et ceci malgré le fait qu'il s'agisse d'un corpus traduit (nous laissons donc de côté la notion de *translationese*, Jonasson 2006).

5 Et ceci malgré l'utilisation de notions théoriques usuelles en sémantique lexicale pour l'étude de ce nom, démarche à laquelle nous ne pouvons que renvoyer ici : Tauveron (2012, à paraître) et Tauveron (en préparation).

De la cooccurrence généralisée à la variation du sens lexical

- (1) Nous organiserons et soutiendrons des *actions* comme, en particulier, le débat sur le besoin d'élaborer un programme législatif pour la mise en œuvre du principe de la reconnaissance mutuelle des décisions judiciaires en matière criminelle. (traduit par *åtgärd*)
- (2) Je pense toutefois qu'il est tout à fait évident qu'au moment de prendre peu à peu une décision sur une mesure, on est souvent face à de nombreuses possibilités d'*action* différentes. (traduit par *handling*)

C'est ce fait qui nous incite à étudier le sens du nom *action* dans ce corpus, sans se focaliser sur son cotexte immédiat, mais en élargissant le champ d'observation.

Nous posons ici notre seconde hypothèse, selon laquelle les différences de sens du nom *action* d'une occurrence à l'autre sont explicables si on envisage ce nom dans un contexte suffisamment étendu. Ce faisant, la taille du corpus à traiter devient trop importante pour en faire une analyse manuelle et justifie l'utilisation d'un outil de statistique exploratoire qu'est le graphe de cooccurrence (défini section 2, cf. Figure 1 *infra*). Ce graphe donne ainsi une vision synthétique du lexique de l'ensemble du texte (ou d'un corpus), et de la manière dont il est utilisé. Comme on le verra, cette vision synthétique permet de mettre en évidence des faits sémantiques pertinents qui expliquent la différence de traduction constatée.

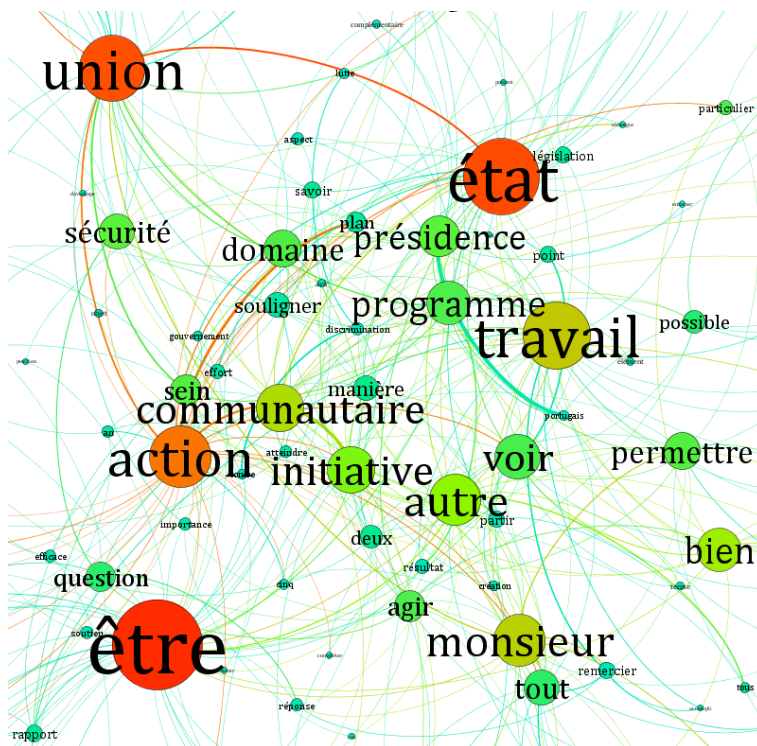


Figure 1. Exemple de graphe de cooccurrence (corpus Europarl)

1.2.2 Constitution du corpus

Le corpus utilisé pour cette étude est un corpus parallèle (Granger 2003, Andersson & Norén 2010) formé à partir d'Europarl, qui fournit au départ 400 occurrences de *action* avec un cotexte élargi (environ 60 mots de part et d'autre). Le corpus est trié manuellement en fonction de l'unité qui traduit le nom *action*. Le corpus est ensuite réparti en deux parties : l'une contenant les occurrences traduites par *handling* (dorénavant, corpus AH), et l'autre contenant celles traduites par *åtgärd* (dorénavant, corpus AA)⁶.

6 Nous laissons de côté les 20 traductions minoritaires attestées.

1.2.3 Traitement informatique

Le texte est ensuite lemmatisé par un programme Perl qui exploite le dictionnaire fourni par l'ABU⁷. Puis vient l'étape de constitution du graphe de cooccurrence à proprement parler, elle aussi effectuée grâce un programme Perl écrit par nos soins, qui fournit la description du graphe dans un format CSV utilisable par le logiciel de traitement de graphes Gephi⁸.

Selon nous, trois questions linguistiques se posent alors.

En premier lieu, la largeur de la fenêtre de cooccurrence utilisée. Nous considérons ici arbitrairement que deux termes sont cooccurents lorsqu'ils sont co-présents dans une fenêtre contenant 7 mots contigus.

En deuxième lieu, le statut des ponctuations fortes délimitant des phrases. Nous avons considéré que les coupures de phrases n'étaient pas des limites suffisamment pertinentes⁹ pour interrompre les liens de cooccurrence.

En troisième lieu, pour des raisons techniques, c'est à l'échelle du mot que porte la relation de cooccurrence. Il nous est encore malheureusement impossible de travailler sur des unités proprement sémantiques, comme l'a fait Valette (2010) en travaillant sur des sèmes.

1.3 Démarche

Nous exposerons dans un premier temps plus largement ce qu'est un graphe de cooccurrence (2.). Nous montrons grâce à cet outil mathématique que la cooccurrence généralisée dispose de deux formes d'organisation complémentaires, que nous envisagerons séparément. En premier lieu (3.), il y a une organisation hiérarchique, qui se manifeste dans le caractère plus ou moins central de certaines unités dans le graphe. Cette centralité est mesurée par le paramètre numérique qu'est la *betweenness*

7 Disponible : <http://abu.cnam.fr/DICO/mots-communs.html>. Si cette procédure n'est pas sans défauts (non-exhaustivité consubstantielle à la nature même du dictionnaire, aucune gestion de l'homonymie), ces dysfonctionnements sont pour la plupart repérés par le lemmatiseur, et laissent la possibilité à une correction humaine par la suite.

8 Nous utilisons Gephi 0.7beta (www.gephi.org). Tous les calculs mentionnés sont effectués par Gephi.

9 Noailly (2002), en particulier pour des retranscriptions d'oral (Benzitoun 2004).

*centrality*¹⁰. En second lieu (4.), il y a une organisation plus démocratique, qui montre que les unités se regroupent en classes du fait de leurs affinités. C'est l'organisation modulaire du graphe.

Le principe sous-jacent à cette étude est que, si *action* est traduit tantôt par *handling* et tantôt par *åtgärd*, c'est notamment pour des raisons touchant au sens global des textes et aux contextes de *action*. Dans le premier temps (3.), nous chercherons des différences de sens à l'échelle globale des deux corpus. Dans le second temps (4.), nous identifierons les lemmes du texte avec lesquels il a des affinités.

2. Graphes de cooccurrence

2.1 Construction et définition du graphe de cooccurrence

Le graphe de cooccurrence représente chaque lemme du texte sous la forme d'un point, et relie deux points si et seulement les deux lemmes correspondants sont cooccurents (Véronis 2004). Chaque lien et chaque nœud sont affectés d'un poids plus important à mesure que le lemme correspondant ou la cooccurrence en question sont plus fréquents. Sur la Figure 1 (cf. *supra*), l'importance des nœuds et des liens est représentée par leur taille. L'ensemble du graphe donne accès à une vue synthétique de toutes les relations de cooccurrence dans un corpus donné. Les relations de cooccurrences établissent ainsi une sorte de topographie, montrant l'organisation du lexique dans le discours.

Au-delà de son caractère ergonomique, le graphe est une structure mathématique, qui dispose de descripteurs définis dont certaines applications ont été bien décrites (Albert & Barabasi 2002).

2.2 La relation de cooccurrence généralisée

La première application d'un tel graphe est qu'il propose une nouvelle représentation de la cooccurrence généralisée telle que l'a décrite Viprey (2006). Grâce à cet outil, on peut mener une étude de la textualité qui fait le lien entre le niveau du texte et le

10 On l'appelle parfois en français *centralité d'intermédiation*. Nous gardons ici l'appellation anglo-saxonne et son abréviation courante dans la littérature (BC).

niveau du lexique. En effet, étudiant la cooccurrence généralisée, nous prétendons parcourir l'ensemble d'un texte donné au travers de la relation de cooccurrence. « [O]n peut voir en effet le texte comme une entité réticulaire et la cooccurrence généralisée comme l'essence de cet entrelacs de mots corrélés, de ce tissu d'associations privilégiées » (Luong *et al.* 2010 : 322). La démarche que nous proposons ici permet de donner une vue synthétique des relations au sein de ce réseau.

Par ailleurs, nous interprétons la cooccurrence – phénomène formel, coprésence matérielle des mots dans le texte – comme une corrélation, c'est-à-dire que nous faisons une interprétation sémantique de cet arrangement formel (Mayaffre 2008). Passer de la sorte du plan de l'expression au plan du contenu est essentiel, et définitoire (d'après Rastier 2011 : 44) de la sémantique reposant sur l'analyse de corpus et la lexicométrie.

2.3 Faits de discours étudiés par la cooccurrence généralisée

2.3.1 La textualité

Nous prenons en premier lieu un point de vue strictement interne sur le texte : ne sont jamais envisagés ici les rapports qu'entretient le texte donné avec les autres textes, ou au travers des genres, discours, etc., phénomènes qui ont tous leur part dans la construction du sens et l'organisation du discours.

Par ailleurs, le graphe de cooccurrence donne une image figée du texte tel qu'il apparaît dans son entier, sans pouvoir envisager quoi que ce soit de son déroulement. Est ainsi laissée de côté ce que Legallois (2006) appelle la « dimension incrémentielle » du texte. C'est-à-dire le fait que celui-ci, sans se réduire à une dimension simplement linéaire ou séquentielle, *avance* par accumulation successive.

2.3.2 Le sens lexical

En montrant les cooccurents de tout lemme du texte, et ceci de manière structurée et qualitative¹¹, le graphe de cooccurrence donne accès aux cooccurents, puis aux cooccurents des cooccurents, et à leurs relations mutuelles et ainsi de suite. Il

11 Cette vue est qualitative au sens où la représentation visuelle montre par exemple, mieux qu'une liste, la plus ou moins grande densité des différentes parties du réseau de cooccurrence, ou le caractère central de tel ou tel mot dans le réseau de cooccurrence.

montre donc la singularité de l'emploi d'un mot donné dans un corpus. Comme nous le verrons, envisager qualitativement le réseau de cooccurrence permet de rendre compte de la variabilité sémantique d'une unité lexicale.

Nous représentons sur la Figure 2 *infra* une conception de l'homonymie, de la polysémie et de la variabilité sémantique héritée de Victorri & Fuchs (1996 : 86 notamment) que nous suivons ici.

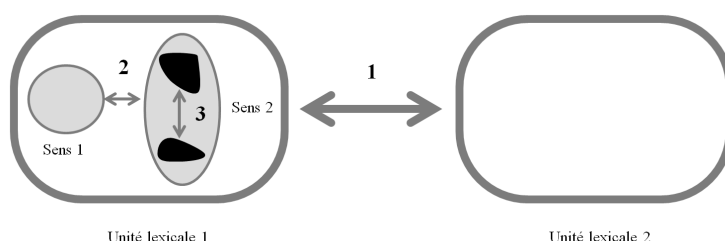


Figure 2. Représentation de la variabilité au sein d'un unique sens lexical

Sur cette figure, la relation 1 est une relation d'homonymie, établie entre deux unités entre lesquelles il n'y a pas de rapport de sens. La polysémie (relation entre les différents sens d'une même unité) est représentée par la relation 2. Elle met en relation *un sens* avec *un autre sens*. Le phénomène que nous étudions ici n'est pas une différence *entre deux sens*, mais une variabilité *au sein d'un même sens*, car nous supposons (avec Victorri & Fuchs, *op. cit.*) qu'un sens n'est pas une entité ponctuelle, mais une région d'un espace. Nous considérerons en effet que, puisque l'intuition linguistique ne perçoit pas *a priori* la différence entre nos occurrences 1 et 2 (section 1.1.1) comme le passage d'un sens à un autre, elles actualisent toutes les deux le même sens en langue du nom *action*. La différence que met en avant la traduction est due au discours. C'est-à-dire que les différences que nous cherchons à observer ne relèvent ni de l'homonymie ou de la polysémie telles que traitées par Véronis (2004). Il n'y a qu'un seul sens en langue derrière l'ensemble des occurrences du nom *action* que nous étudions, mais cet unique sens est susceptible de variation.

Le caractère purement interne au texte de l'analyse à laquelle nous procédons implique donc que nous travaillons sur le sens des mots du texte à partir des seuls indices que sont les autres mots du texte. Ceci ne signifie nullement pour autant que le texte construise à lui seul le sens de ses mots. Comme l'a défendu Kleiber (1997, notamment) contre Kayser (1987, notamment), le mot est doté d'un sens en langue, que le contexte vient préciser, influencer dans un sens ou d'autre. Chaque mot dispose certes d'une inertie sémantique, plus ou moins importante, qui fait qu'il ne peut pas *tout* signifier, et que le contexte ne peut pas *tout*. Mais l'objet d'étude abordé ici est justement cette part de variation créée par le contexte.

3. Organisation hiérarchique de la cooccurrence généralisée

3.1 La *betweenness centrality*

*3.1.1 Définition de la *betweenness centrality**

La *betweenness centrality* permet de quantifier la centralité d'un nœud dans un graphe – ce que nous interprétons comme un indicateur de saillance. Parmi les mesures disponibles dans la littérature pour chiffrer cette importance, notre choix s'est porté sur celle-ci pour trois raisons :

1. elle reflète l'intuition : les unités qui semblent à l'œil nu être plus importantes dans le graphe ont une BC plus élevée, contrairement aux résultats fournis par les autres mesures de centralité (Wasserman & Faust 1994 : 215 *sq.*) ;
2. elle est en général assez bien corrélée à la fréquence, sans s'y réduire : les lemmes les plus fréquents du texte ont souvent la BC la plus importante. La BC amplifie cependant des différences ténues entre les lemmes de fréquence analogue, et éclaire donc le classement.
3. Par définition, elle renvoie à une forme d'organisation des unités pertinente pour notre problématique : les unités ayant la BC la plus élevée ont à la fois un rôle organisateur dans le graphe, et constituent un intermédiaire entre les différentes régions de celui-ci (Vergès & Bouriche 2001 : 69). Dans le cas du graphe de co-

occurrence, les unités dotées de la BC la plus élevée apparaissent comme des dénominateurs communs à différents contenus du texte.

Ce dernier fait découle directement de la définition de la BC. On obtient en effet la BC d'un nœud N donné de la manière suivante (Anthonisse 1971). Etant donnés deux nœuds quelconques, on recherche tous les chemins qui relient ces deux nœuds, parmi lesquels on ne considère que celui ou ceux de longueur minimale. On calcule la probabilité que le plus court chemin entre ces deux nœuds passe par N (*i.e* la proportion de ces plus courts chemins qui passent par le nœud N). Ce calcul est fait sur toute paire de nœuds dans le graphe, et la BC est la somme de toutes ces probabilités. Dans le cas particulier du graphe de cooccurrence, Xie (2005) a montré que la BC indiquait les unités sémantiquement les plus saillantes du texte¹².

3.1.2 L'organisation hiérarchique du graphe

Un examen de la BC permet donc de montrer l'organisation hiérarchique du lexique du texte : les unités dotées de la BC la plus élevée sont les plus saillantes, une BC moins élevée indique une saillance moindre, et ainsi de suite. L'organisation mutuelle entre ces unités saillantes est montrée par les liens du graphe.

3.2 Comparaison entre les deux corpus

3.2.1 Les unités les plus saillantes

Les classements des unités les plus centrales dans les deux graphes sont similaires (Tableau 1). Cela montre que les deux corpus se ressemblent, et traitent, dans les grandes lignes, des mêmes thématiques.

12 En l'espèce, les noms dotés de la BC la plus élevée dans un texte ont une d'autant plus grande probabilité d'apparaître dans les résumés de ces textes faits par les humains.

Tableau 1. BC des unités les plus centrales dans les deux corpus

Rang	Corpus AA		Corpus AH	
	Lemme	BC	Lemme	BC
1	programme	1440	commission	2022
2	commission	1345	social	1061
3	européen	833	européen	1042
4	politique	685	politique	465
5	président	483	femme	435
6	conseil	406	président	361
7	action	349	état	248
8	développement	298	question	234
9	énergie	245	droit	233

Les deux emplois de *action* que nous observons sont donc également similaires. Est donc confirmée notre hypothèse de départ, à savoir qu'on ne peut pas parler de polysémie au sujet de ces deux emplois.

Nous allons nous attarder sur la différence la plus significative entre les deux corpus : le rôle important du nom *programme* dans le corpus AA.

3.2.2 Le nom *programme* dans le corpus AA

La Figure 3 (*infra*) permet de voir que ce nom joue un rôle central dans ce corpus, aux côtés de *commission*, *européen* et *politique*, où il est relié à de nombreuses unités différentes. C'est par son intermédiaire qu'un grand nombre d'unités, d'importance moyenne ou secondaire, sont reliées au reste du graphe, ainsi *action*, *initiative*, *communautaire*, *énergie*, *culture*, *objectif*. Dans le corpus AH, il joue en revanche un rôle périphérique, faisant le lien entre des thématiques plus importantes que lui (*commission*, *social*), et entre celles-ci et des unités moins importantes (*action*, *élaboration*).

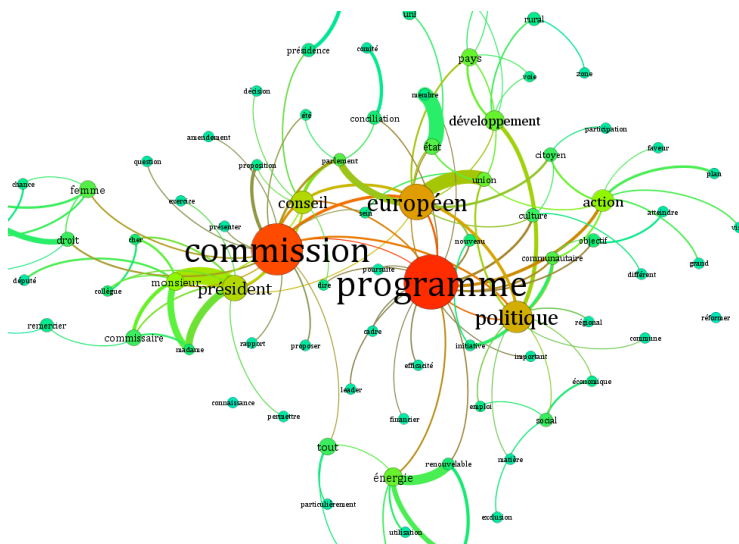


Figure 3. Détail du graphe du corpus AA

Programme joue donc bien un double rôle d'organisateur et d'intermédiaire entre différents contenus. Cette observation permet de supposer qu'un certain nombre de thématiques sont envisagées, dans ce corpus, sous un angle programmatique ou virtuel. Un retour sur les données textuelles le confirme :

- (3) Il apparaît, en outre, très clairement qu'un certain nombre d'Etats membres ne sont guère enclins à vouloir réduire la capacité de leurs flottes de pêche conformément aux objectifs et décisions des programmes. Dans le prolongement d'un renforcement des contrôles, la Commission est invitée à mettre en œuvre des actions – et, le cas échéant, à renforcer les actions déjà mises en œuvre [...]. (Corpus AA, occurrence n° 107)

Dans cette occurrence, le nom *programme* contribue à construire une dimension virtuelle dans le discours, également manifestée par être *enclin à* (avec l'atténuateur *ne ... guère*) et *être invité à*. On retrouve un caractère analogue dans cette occurrence :

- (4) [Le Programme Culture 2000] encourage la mobilité, et il ouvre les portes de la culture aux personnes défavorisées [...]. Mon seul regret est que nous ne disposions pas d'un

financement suffisant, dans la promotion de ce programme, pour atteindre nos aspirations et garantir notre capacité de réalisation. Nous disposerons ainsi d'un outil tout à fait apte à développer, au cours des cinq prochaines années, une action claire, bien structurée et, j'en suis sûre, fructueuse, [...]. (Corpus AA, occurrence n° 180)

La virtualité est ici dénotée par des verbes et des locutions qui renvoient à des actions non effectuées (*encourager, ouvrir les portes de la culture, regret*, la référence à un financement insuffisant, *promotion, atteindre des aspirations, garantir, apte à faire*), en plus de l'emploi du futur (*disposerons*) et du modalisateur *j'en suis sûr* qui vont tous deux dans le sens d'une virtualité.

3.3 Conclusions sur l'organisation hiérarchique

3.3.1 Organisation du lexique dans les textes

L'étude de la cooccurrence généralisée sous forme de graphe a permis de montrer une organisation hiérarchique du lexique dans les textes. En effet, le calcul de la BC permet de déceler les lemmes ayant une position centrale dans la relation de cooccurrence généralisée, et que d'autres ont une position marginale. Malgré cette hiérarchie, le graphe n'est pas pour autant disposé en arbre : il n'y a aucune unité principale dont dépendraient les autres. Ceci s'explique par le fait que la hiérarchie se superpose à d'autres formes d'organisation que nous exploiterons par la suite.

3.3.2 Visualisation des notions

Cette représentation de la cooccurrence généralisée permet de se faire une idée des lemmes les plus saillants du texte. Comme on l'a vu, un tel outil permet de montrer – comme le fait une liste de fréquences – des différences globales entre deux corpus.

Cependant, la valeur ajoutée du graphe de cooccurrence est qu'il représente la manière dont ces unités s'organisent et se regroupent pour former des contenus sémantiques de fond. Nous pensons que l'on peut ainsi aborder une représentation des « notions » au sens de Culioli :

Il y a place [...] pour des chaînes d'associations sémantiques où l'on a des « grappes » de

propriétés établies par l'expérience, stockées et élaborées sous des formes diverses [...]. Cette ramification de propriétés qui s'organisent les unes par rapport aux autres [...] établit ce que j'appelle un domaine notionnel. C'est une représentation sans matérialité, ou plutôt dont la matérialité est inaccessible au linguiste. Les notions ne correspondent donc pas directement à des items lexicaux. (Culioli 1991 : 10)

4. Organisation modulaire de la cooccurrence généralisée

4.1 La notion de classe de modularité

4.1.1 Une forme d'organisation en classes

On a vu comment le paramètre numérique qu'est la BC permet de déceler une organisation hiérarchique dans la cooccurrence généralisée. Cette organisation ne permet cependant pas une étude du comportement d'une unité considérée *a priori*, comme le nom *action* dans notre cas. Nous allons nous focaliser ici sur une autre forme d'organisation de la cooccurrence généralisée, non hiérarchique, et plus locale.

L'étude de la modularité d'un graphe consiste à répartir ses nœuds en classes, chacune étant soudée par des liens de cooccurrence. Par construction, les nœuds situés dans une même classe sont bien reliés entre eux, et peu reliés aux nœuds appartenant à une autre classe¹³. Sur la Figure 4, chaque classe est associée à une couleur.

Cependant, l'intérêt de cette notion réside dans sa subtilité : les classes sont poreuses et ponctuellement reliées les unes des autres, car tous les nœuds d'un graphe sont reliés au moins indirectement.

¹³ Parmi tous les algorithmes décrits dans la littérature pour obtenir ce genre de résultats, Gephi utilise celui mis au point par Blondel *et al.*, 2008.

Les nœuds et les liens d'un graphe de cooccurrence donnent en effet une image des unités d'information d'un texte. Les répartir en différentes classes, et montrer la manière dont elles sont hiérarchisées et reliées est une véritable identification du contenu thématique, ou de la structuration thématique du texte (Wilson 1968, Hjørland 2001). C'est précisément ce que permet l'étude des classes de modularité.

Pour répondre à notre problématique sur le nom *action*, nous allons nous focaliser, dans chacun de nos deux corpus, sur la classe de modularité dont fait partie ce nom, pour relever les termes avec lesquels il est en relation de manière privilégiée.

4.2 Le placement du nom *action* dans les deux corpus

*4.2.1 Le nom *action* dans le corpus AH*

Dans ce corpus, on peut mettre en avant deux types d'unités de langue parmi celles avec lesquelles *action* noue des relations privilégiées :

1. d'une part des noms d'entités, qui dans le corpus sont des noms d'agents, ainsi *pays*, *état* et *membre* ;
2. d'autre part des unités faisant référence à la planification des actions : *directive*, *viser*.

En particulier, aucun terme ne renvoie ni à l'effectuation des actions, ni aux résultats obtenus.

Cette occurrence apparaît par exemple comme typique de cette partie du corpus :

- (5) En ce qui concerne les premières, la Commission nous a envoyé une série de propositions, contenues dans deux directives et un plan d'action, qui cherchent à appliquer la décision suivante prise par les chefs d'Etats et de gouvernements à Tampere : « assurer un traitement équitable aux ressortissants de pays tiers qui résident légalement sur le territoire de ses Etats membres et une politique plus énergique en matière d'intégration qui devrait avoir pour ambition de leur offrir des droits et obligations comparables à ceux des citoyens de l'Union ». (Corpus AH, occurrence 394)

*4.2.2 Le nom *action* dans le corpus AA*

Dans le corpus AA, *action* fait partie d'une classe de modularité bien différente. Si on rencontre un certain nombre d'unités

communes avec la classe de modularité précédente, on remarque *a contrario* 4 types d'unités de langue qui font contraste :

1. des unités qui indiquent que l'action est dirigée vers quelque chose : *contre, envers, destinée* ;
2. des noms concrets ou abstraits qui, dans le contexte, dénotent des thématiques ou des domaines sur lesquels porte l'action : *pauvreté, chômage, exclusion, criminalité* (problèmes que règle l'action) et *égalité* (objectif de l'action) ;
3. l'adjectif *concret*, qui dans ce cas indique que l'action a des conséquences ;
4. des noms d'actions, qui, dans le contexte, apparaissent comme servant à dénoter les programmes suivis par l'action : *protection, prévention*.

Dans le corpus AH, l'action est envisagée sous l'angle de ses conséquences et de son effectuation. A ce titre, cette occurrence apparaît comme typique :

- (6) Des propositions concrètes tant pour la législation que pour des plans d'action qui concernent la lutte contre les discriminations, et qui concernent l'établissement de libertés et droits sociaux. J'en viens maintenant à l'initiative EQUAL laquelle est bien sûr basée sur l'article 13. L'initiative EQUAL concerne toutes les formes de discriminations sur la base de l'article 13, c'est-à-dire la lutte contre les discriminations fondées sur le sexe, la race, l'âge, le handicap. (Corpus AA, occurrence 231)

4.3 La mise en saillance de zones particulières du sens lexical

La comparaison des classes de modularité de *action* a donc permis de mettre en évidence une différence sémantique entre les emplois de ce nom dans les deux corpus. On a ainsi pu montrer que les deux traductions de *action* en suédois correspondaient bien à une différence sémantique en français. On ne peut parler de polysémie entre ces deux emplois : il faut supposer qu'ils sont sous-tendus par un unique schéma commun (Lakoff 1987), mis en œuvre de deux manières distinctes (Ruiz de Mendoza 1998). Nous mettons cette différence de mise en œuvre sur le compte d'une différence de saillance dans la représentation du même référent.

Les observations des deux paragraphes précédents amènent à adopter une représentation simple du schéma sous-tendu par le nom *action* (à la Langacker 1987) : en plus d'un processus (qui constitue le référent du nom à proprement parler), on fait figurer ce qui se trouve en amont de lui (les causes, au nombre desquelles l'agent), et ce qui se trouve en aval (les résultats de l'action). Au vu des analyses faites dans les paragraphes précédents, il apparaît que les deux emplois de *action* (dans AH et dans AA) se caractérisent, l'un par une mise en saillance de l'amont de l'action, l'autre par une mise en saillance de l'aval :

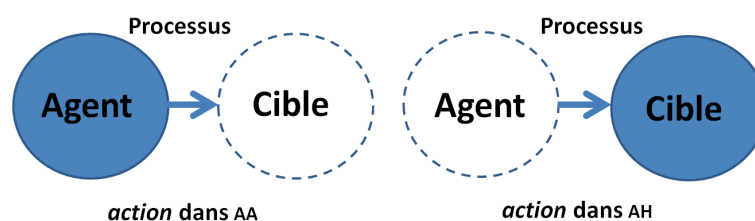


Figure 5. Mises en saillance distinctes sur le même schéma pour les deux emplois du nom *action*

5. Conclusion : la sémantique de la cooccurrence généralisée

5.1 Organisation du lexique dans les énoncés

L'étude faite sur la BC d'une part, et sur les classes de modularité d'autre part a montré la présence de deux organisations complémentaires dans la cooccurrence généralisée. En premier lieu, il apparaît que le lexique des textes est organisé hiérarchiquement : des différences de saillance entre unités lexicales donnent un rôle prééminent à certaines d'entre elles. Ces unités prééminentes ont une position centrale dans le réseau, ce qui leur donne une certaine influence sémantique sur les autres. En second lieu, cette hiérarchie se double d'une organisation plus démocratique, qui regroupe les lemmes en classes selon leur tendances à être co-occurents les uns avec les autres. Cette organisation modulaire scinde le lexique du texte en différentes parties, que l'on peut interpréter comme des thèmes.

En plagiant les propos d'Adam (2004 : 35) sur les propositions, nous avons donc pu montrer différents « agencements complexes mais non anarchiques » de mots dans les textes.

5.2 Variabilité du sens en discours : la notion de micro-sens

Passant de la cooccurrence à la corrélation, nous avons pu faire une interprétation sémantique des faits observés. Nous avons pu en effet montrer que les deux traductions en suédois du nom *action* correspondaient en français à deux emplois, certes distincts, d'un même sens. Il ne s'agit pas là d'un paradoxe, mais d'une possibilité fondamentale du sens linguistique, celle d'une « invariabilité et variabilité non finie, ouverte, du sens d'un même mot en emploi » (Durafour 2005 : 85).

Dans le cas précis du nom *action*, la variabilité que nous avons constatée est peut-être à représenter sous forme de *micro-sens* (Croft & Cruse 2004 : 127). Les deux valeurs du nom *action* sont en effet des réalités sémantiques distinctes, qui ne peuvent s'actualiser sur la même occurrence, et pourtant associées au même signifiant. De surcroît, malgré leur incompatibilité, elles peuvent passer pour des valeurs sous-ordonnées à une valeur plus générale, également dénotée par le nom *action* (dans laquelle ni l'agent ni la cible ne sont mis en saillance). La notion de micro-sens apparaît comme tentante, dans la mesure où elle rend compte à la fois de la similarité et la différence entre deux valeurs rencontrées en discours, puisqu'elles sont à considérer comme deux cohyponymes d'une même valeur hyperonymique.

5.3 Le pouvoir de l'actualisation en discours

Les phénomènes mis en avant ici (organisation du lexique, variabilité du sens) sont – il est nécessaire de le souligner – des réalités construites par le discours. En particulier, les organisations hiérarchique et modulaire ne sont pertinentes que dans un texte donné, et ne sont pas à relier à des réalités de la langue : c'est l'actualisation en discours qui met en saillance telle unité lexicale, ou qui place deux unités dans la même classe.

C'est dans la même perspective qu'il faut souligner que la différence d'emplois mise en évidence ici n'est pas un signe de polysémie. Nous nous rangeons en effet à l'avis de Kleiber (2005 : 69) pour lequel « ce n'est que si le changement est

conventionnalisé ou préconstruit qu'il y a polysémie ». Laissant de côté les conventions ou les préconstructions de la langue, notre étude a porté sur la timide mais bien existante variabilité du sens construite par l'actualisation du mot en discours.

Références bibliographiques

- Adam J.-M. (2004). *Linguistique textuelle. Des genres de discours aux textes*. Paris : Nathan.
- Albert R. & Barabasi L. (2002). « Statistical mechanics of complex networks », *Review of Modern Physics* 74, 1 : 47-97.
- Andersson C. & Norén C. (2010). « Comparer la finalité dans le débat parlementaire : l'apport du corpus bilingue C-ParlEur », *Cahiers Sens public* 13-14 : 35-53.
- Anthonisse J.M. (1971). « The rush in a directed graph », *Stichting Mathematisch Centrum. Mathematische Besliskunde* 9, 71 : 1-10.
- Benzitoun C. (2004). « L'annotation syntaxique de corpus oraux constitue-t-elle un problème spécifique ? », Actes de la conférence RECITAL, 19-22 avril, Fès, 13-22.
- Blondel V., Guillaume J.-L., Lambiotte R. & Lefebvre E. (2008). « Fast unfolding of communities in large network », *Journal of Statistical Mechanics : Theory and Experiment*, P10008.
- Bonneau J., Anderson J., Stajano F. & Anderson R. (2009). « Eight Friends Are Enough : Social Graph Approximation via Public Listings », SNS '09, Nuremberg.
- Croft W. & Cruse D.A. (2004). *Cognitive Linguistics*. Cambridge : Cambridge University Press.
- Culioli A. (1991). « Structuration d'une notion et typologie lexicale. A propos de la distinction dense, discret, compact », in Culioli A. (1999). *Pour une linguistique de l'énonciation*, tome 3. Paris : Ophrys, 9-17.
- Durafour J.-P. (2005). « De la double référence du langage en sémantique génétique », in Murguia A. (éd.), *Sens et*

références : mélanges Georges Kleiber. Tuebingen : Gunter Narr Verlag, 83-120.

- Fortunato S. & Castellano C. (2007). « Community Structure in Graphs », Disponible : <http://lanl.arxiv.org/abs/0712.2716>.
- Granger S. (2003). « The corpus approach : a common way forward for Contrastive Linguistics and Translation Studies ? », in Granger S., Lerot J. & Petch-Tyson S. (eds), *Corpus-based Approaches to Contrastive Linguistics and Translation Studies*. Amsterdam : Rodopi, 17-29.
- Hjørland B. (2001). « Towards a Theory of Aboutness, Subject, Topicality, Theme, Domain, Field, Content... and Relevance », *Journal of the American Society for Information Science and Technology* 52, 9 : 774-778.
- Jonasson K. (2006). « Démonstratifs indéfinis en suédois et leurs correspondants en français », *Langue française* 152 : 24-38.
- Kasper C. & Voelkl B. (2009). « A social network analysis of primate groups », *Primates* 50 : 343-356.
- Kayser D. (1987). « Une sémantique qui n'a pas de sens », *Langages* 87 : 33-45.
- Kleiber G. (1997). « Sens, référence et existence : que faire de l'extra-linguistique ? », *Langages* 127 : 9-37.
- Kleiber G. (2005). « Quand y a-t-il sens multiple ? Le critère référentiel en question », in Soutet O. (dir.), *La polysémie*. Paris : Presses de l'Université Paris-Sorbonne, 51-73.
- Koehn P. (2005). « Europarl : A Parallel Corpus for Statistical Machine Translation », MT Summit 2005, disponible : <http://homepages.inf.ed.ac.uk/pkoehn/publications/europarl-mtsummit05.pdf>.
- Lakoff G. (1987). *Women, Fire and Dangerous Things. What Categories Reveal about the Mind*. Chicago University Press.
- Langacker R. (1987). *Foundations of Cognitive Grammar*, vol. 1. Stanford University Press.

- Legallois D. (2006). « Présentation générale. Le texte et le problème de son et ses unités : propositions pour une déclinaison », *Langages* 163 : 3-9.
- Luong X., Brunet E., Longrée D., Mayaffre D., Mellet S. & Poudat C. (2010). « La cooccurrence, une relation asymétrique ? », in Bolasco S., Chiari I. & Giuliano L., *Statistical Analysis of Textual Data. Proceedings of 10th International Conference Journées d'Analyse statistique des Données Textuelles*, 321-331.
- Lusseau D., Whitehead H. & Gero S. (2008). « Incorporating uncertainty into the study of animal social networks », *Animal Behaviour* 75, 5 : 1809-1815.
- Mayaffre D. (2008). « De l'occurrence à l'isotopie. Les co-occurrences en lexicométrie », *Syntaxe et sémantique* 9 : 53-72.
- Noailly M. (2002). « L'ajout après un point n'est-il qu'un simple artifice graphique ? », in Authier-Revuz J. & Lala M.-C. (éd.), *Figures d'ajouts. Phrase, texte, écriture*. Paris : Presses Sorbonne Nouvelle, 133-145.
- Paranyushkin D. (2010). « Text network analysis », *Performing Arts Forum*. Disponible : <http://noduslabs.com/research/pathways-meaning-circulation/>.
- Rastier F. (2011). *La Mesure et le grain*. Paris : Champion.
- Ruiz De Mendoza F. J. (1998). « On the nature of blending as a cognitive phenomenon », *Journal of Pragmatics* 30 : 259-274.
- Tauveron M. (2012, à paraître). « Une ontologie multidomaine des événements. Noms d'événements comme noms d'éten-dues », *Syntaxe & Sémantique* 13.
- Tauveron M. (en préparation). Sémantique et cognition : les noms généraux d'entités du 2^{ème} ordre dans une perspective contrastive français-suédois. Thèse de doctorat à l'Université de Strasbourg.
- Valette M. (2010). « Propositions pour une lexicologie textuelle », *Zeitschrift für Französische Sprache und Literatur* 37 : 171-188.

- Vergès P. & Bouriche B. (2001). « L'analyse des données par les graphes de similitude ». Disponible : <http://www.scienceshumaines.com/textesInedits/Bouriche.pdf>.
- Véronis J. (2004). « HyperLex : Lexical Cartography for Information Retrieval », *Computer Speech & Language* 18, 3 : 223-252.
- Victorri B. & Fuchs C. (1996). *La polysémie : construction dynamique du sens*. Paris : Hermès.
- Viprey J.-M. (2006). « Structure non-séquentielle des textes », *Langages* 163 : 71-85.
- Wasserman S. & Faust K. (1994). *Social Network Analysis*. Cambridge : Cambridge University Press.
- Wilson P. (1968). *Two kinds of power. An essay on bibliographical control*. Berkeley : University of California Press.
- Xie Z. (2005). « Centrality Measures in Text Mining : Prediction of Noun Phrases that Appear in Abstracts », ACL'05, Proceedings of the Student Research Workshop, Ann Arbor.